Unifying DNA methylation-based in silico cell-type deconvolution with *methyldeconv*

Alexander Dietrich[1],* Lina-Liv Willruth[1], Korbinian Pürckhauer[1], Carlos Oltmanns[2,3,4], Moana Witte[2,3,4], Sebastian Klein[2,3,4,5] Anke RM Kraft[2,3,4,6,7], Markus Cornberg[2,3,4,6,7] and Markus List[1,8] *

[1] Data Science in Systems Biology, TUM School of Life Sciences, Technical University Munich, Munich, Germany
[2] Centre for Individualised Infection Medicine (CiiM), a joint venture between the Helmholtz Centre for Infection Research (HZI) and Hannover Medical School(MHH), Hannover, Germany
[3] Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School (MHH), Hannover, Germany
[4] TWINCORE, a joint venture between the Helmholtz-Centre for Infection Research (HZI) and the Hannover Medical School (MHH), Hannover, Germany
[5] CAIMed – Center for AI in Medicine, Joint Venture of Leibniz University Hannover and Hannover Medical School, Hannover, Germany
[6] Cluster of Excellence Resolving Infection Susceptibility (RESIST; EXC 2155), Hannover Medical School, Hannover, Germany
[7] German Centre for Infection Research (DZIF), Partner Site Hannover-Braunschweig, Braunschweig, Germany
[8] Munich Data Science Institute (MDSI), Technical University of Munich, 85748 Garching, Germany
* Corresponding authors

Abstract

Summary

Cell-type deconvolution is widely applied to gene expression and DNA methylation data, but access to methods for the latter remains limited. We introduce *methyldeconv*, a new package that improves access to DNA methylation-based deconvolution methods, and compare their estimates to those from gene expression and experimental ground truth data.

Availability and Implementation

MethylDeconv is available at https://github.com/omnideconv/methyldeconv. The scripts to conduct the analysis steps and figures can be found at https://github.com/omnideconv/methyldeconv_analysis.
Matched EPIC array data, RNA-seq data, and flow cytometry values are available at https://doi.org/10.6084/m9.figshare.28563854.v1.

Contact

Markus List <markus.list@tum.de>; Alexander Dietrich <alex.dietrich@tum.de>

Supplementary information

Additional figures and tables are available online.

Introduction

The human body consists of many cell types whose differentiation is characterized by changes in epigenomic and transcriptomic profiles. The exact composition of cell types in samples can provide valuable insights into, e.g., disease subtype or treatment response [1]. Experimental techniques to measure cell-type composition include flow cytometry and immunohistochemistry (IHC), which both are limited in their throughput and the number of possible cell types being profiled [2]. The most cost-effective and most widely used assays for studying the epigenome or transcriptome, RNA sequencing (RNA-seq) [3] and Illumina arrays (EPIC or 450k), respectively, produce bulk profiles where the exact cellular composition is unknown. A wide variety of deconvolution methods have been proposed for both gene expression and DNA methylation (DNAm) data. These use cell-type specific gene expression or DNAm markers to inform the deconvolution of bulk samples. Specifically, these methods estimate the cell-type composition from bulk mixtures by solving linear equations [4]. Most deconvolution methods use regression-based algorithms to solve the deconvolution problem, such as non-negative least squares regression (NNLS) [5] or nu-support vector regression (nu-SVR) [6]. Other approaches include optimization and projection methods such as linear constrained projection solved with quadratic programming, where optimal projections under given constraints are reached [7], as well as generative and latent variable models [8]. As no single method outperforms all others, running several approaches in parallel is often necessary. To this end, we have previously developed the *immunedeconv* package [9] for first-generation transcriptomics deconvolution methods that rely on pre-built gene signatures and *omnideconv [10]* for second-generation methods that dynamically learn cell-type specific gene signatures from single-cell expression data. However, for DNAm-based methods, a similar package does not exist. While deconvolution methods for gene expression [9,10] and DNAm-based methods [11,12] have been extensively benchmarked, it is unclear which of the two data types offers more precise cell-type estimates. While selected methods such as *EpiDISH* [11] and ESTIMATE [13] have been compared previously [14], a dedicated performance comparison of methods and their signatures against a gold standard is currently lacking.

In this study, we present *methyldeconv*, a novel R package that provides uniform access to five reference-based deconvolution methods using DNAm array data: *EpiDISH* [11], the Houseman method [7], *MethAtlas* [5], *methylCC* [8] and *MethylResolver* [15]. To showcase the use of *methyldeconv*, we created a unique set of 59 peripheral blood mononuclear cell (PBMC) samples with matched RNA-seq, EPIC array, and FACS profiles. This data allows us to compare the performance of DNAm-based methods included in *methyldeconv* to gene expression-based methods included in the *immunedeconv* in relation to flow cytometry-derived ground truth estimates.

Materials and Methods

Data Processing

Raw *fastq* files from the RNA sequencing have been processed using the nf-core rnaseq pipeline version 3.10.1 [16,17]. The main steps include a manual inspection of quality using *FastQC*, followed by pseudo-alignment and gene quantification using *Salmon* [18]. Next to

raw gene expression counts, this method generates transcript per million (TPM) counts, which are normalized by gene length and library size and were used as input for deconvolution methods.
DNAm data by Illumina EPIC arrays generated raw *idat* files that have been processed with *RnBeads* 2.0 [19] using the default settings during preprocessing, which removes methylation sites with too many missing values, cross-reactive probes, or those that overlap with known single-nucleotide polymorphisms (SPNs). Additionally, we removed all CpGs with NA values in any sample. The final dataset consisted of 724,058 CpGs. We removed CpGs located on sex chromosomes as they would confound the analysis.

methyldeconv implementation

We implemented *methyldeconv* as an R package that uses a *methylSet [20]* object containing the matrices for methylated and unmethylated probes as input. Some methods can also directly be applied to a beta matrix, while others need this additional level of information. Further, each method can easily be executed using `deconvolute()`, where the method parameter specifies which method should be executed. Additionally, we transform the output of each method to a unified style with samples in rows and estimated cell-type fractions in columns.
To directly access all parameters of the original method, we additionally provide functions called `run_<methodname>()`. Each parameter is initially set to the method's default values. The function will also return the original unchanged output of each method.
Further, we provide a function to execute several selected methods together and return their average results, which can improve the robustness of deconvolution estimates.
Both *MethylCC* and the Houseman method can differentiate between the 450k and EPIC array types, a functionality that we also exposed to the user of *methyldeconv*.

Deconvolution

Deconvolution methods using DNAm data as input were used with default parameters as implemented in methyldeconv. Exceptions to this are *MethylResolver*, with a fixed alpha parameter of 1, *EpiDISH* with fixing the used reference to ‘blood’ and the number of iterations to 100, and *MethAtlas* to only use the part of its signature matrix that was designed to estimate immune cells from EPIC data. The input data for deconvolution was not normalized. Similarly, methods from the immunedeconv package were used with default parameters, with *EPIC* and *quanTIseq* set to expect non-tumor and non-array data. For all three selected methods, we applied TPM normalization of samples, as specified by the authors of the methods.
Following deconvolution, cell types were relabelled to allow comparisons between methods that either differed in the spelling of cell types or provided different annotation resolutions. In the latter case, estimates were summed up to the next ‘higher’ matching cell type (Supplementary Table 1).

Comparison of signature matrices

We compared the genomic overlap of features in signatures of eight deconvolution methods. Signature matrices were extracted from the Github repositories or the respective package data. *MethylCC* uses differentially methylated regions (DMRs) instead of methylation sites as its signature features, which are calculated from the *FlowSorted.Blood.450k* whole blood

reference methylomes for six cell types [21,22] using the `find_dmrs()` function. In the following, the CpGs that are part of these DMRs are used for *methylCC*.
To compare the genomic locations of all features on the same reference genome coordinates, we used the `liftOver()` function of the *rtracklayer* package [23,24] in R to lift over all coordinates to the hg38 genome. This affected all DNAm-based methods, as the Illumina manifest files in which coordinates of 450k and EPIC probes are only available for hg19. The coordinates of genes in the reference matrices of *quanTIseq [25]*, *EPIC [26]*, and *CIBERSORT* [6] were taken from the hg38 reference genome; this includes all exon, intron, and promoter regions of a gene. A small number of genes (2, 0, and 34, respectively) could not be retrieved from this annotation and have been left out for this analysis step. The number of overlapping base pairs between CpGs and genes has been calculated using the `make_comb_mat()` function in the *ComplexHeatmap* package [27].

Application: DNAm vs gene-expression-based deconvolution

To illustrate the usability of *methyldeconv*, we generated sample-matched readouts from chronic Hepatitis C virus-infected patients for DNAm using Illumina EPIC arrays, gene expression using RNA-seq (Supplementary Figure 1) as well as proportions of selected immune cells using flow cytometry. With this unique dataset, we could compare the results of reference-based cell-type deconvolution methods implemented in *methyldeconv* to the matched set of ground-truth flow cytometric data and additionally evaluated their performance against gene-expression-based deconvolution methods included in *immunedeconv*.
We selected five methods that can estimate immune cell proportions from DNAm data to be included in *methyldeconv* and as a base of our comparison: *MethAtlas* [5], *methylCC* [8], *MethylResolver* [15], *EpiDiSH* [11] and the Houseman method [7]. All methods are reference-based, meaning they include a cell-type specific reference based on CpGs or - in the case of *methylCC* - DMRs based on prior knowledge. Similar types of methods exist for gene expression measurements, such as from RNA-seq experiments, and among the currently best-performing methods are *quanTIseq* [25], *EPIC* [26], and *CIBERSORT* [6], all available via the *immunedeconv* package [9]. We selected those three methods as a set of comparisons against the previously mentioned DNAm-based methods. As not all references are designed for the same sets of cell types, we only selected a few representative immune cells that are estimated by almost every method (Supplementary Table 1).

We evaluated the performance using our ground truth from flow cytometric analysis with two metrics, Pearson's R and root mean square error (RMSE). All five DNAm-based methods perform similarly globally (Figure 1A). On a cell-type level, *MethAtlas* struggles with a correct estimation of CD8+ T cells, where it only reaches an RMSE of 0.2 due to overestimating this cell type in all samples; conversely, the same method underestimates CD4+ T cells. Similarly, while *EpiDISH* could perform well globally, as previously shown [12], it displays the weakest performance for NK cells. Using *methyldeconv,* we additionally averaged the estimates of the five included methods, which yields robust results globally but is outperformed on the level of individual cell types.

Comparing the DNAm-based methods to those using gene-expression data (Figure 1B, Supplementary Figure 2), we could spot systematic differences, as only *quanTIseq* could match the performances of the *methyldeconv*-included methods. *EPIC* and *CIBERSORT*, to a lesser extent, significantly overestimate monocytes and NK cells while underestimating CD4+ T cells. These issues result in higher global RMSE values for both methods and even a negative correlation for *EPIC* compared to any DNAm-based method in *methyldeconv*.

As all eight selected methods base their deconvolution algorithm on different signature matrices, we further looked into the genomic overlaps of features included in these signatures (Supplementary Figure 3). Strikingly, only a few CpGs are shared between the DNAm-based deconvolution methods, in the case of MethylCC, not a single position of its cell-type-specific DMRs is covered by any other gene or CpG. As all methods performed similarly, this low overlap shows the variety of potential cell-type specific methylation sites in the human genome. Conversely, genes are more commonly shared between signature matrices, although most are unique to a single method. Interestingly, almost no overlaps are present between the two data types, hinting at different layers of cell-type specificity in these data types.

Conclusion

Despite the increasing popularity of single-cell techniques, bulk profiling is still common for large-scale projects due to the low costs and well-established workflows. Moreover, single-cell profiling is rarely used to study DNA methylation. Cell-type deconvolution methods overcome a limitation of bulk profiling by robustly estimating cell-type fractions. Here, we describe a novel R package *methyldeconv* that offers access to five cell-type deconvolution methods tailored for DNAm data. We compared the performance of these methods to RNA-seq based cell-type deconvolution methods and show that they perform better in terms of RMSE and Pearson correlation compared to most gene-expression-based methods. With *methyldeconv*, we improve the accessibility of these methods and offer additional functionalities, such as the aggregated cell-type estimates of selected methods and streamlined visualization methods.

Figure 1: A: Pearson correlation of cell-type fractions predicted by deconvolution methods versus ground truth in 59 samples from bulk DNAm (columns 1-6) and RNA-seq (columns 6-8). The ground truth is derived from flow cytometry. The column labeled 'combined' represents the averaged cell-type fractions across all five DNAm-based deconvolution methods. B: The heatmap shows the corresponding RMSE values for the same samples and ground truth settings. A 'global' RMSE value is also calculated considering the predictions across all cell types. A heatmap with Pearson correlation is shown in Supplementary Figure 2.

Acknowledgments & Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the CompLS research and funding concept [031L0294 (NetfLID)].

Authorship Contributions

AD and ML devised the concept for methyldeconv. AD and KP created the initial version of methylDeconv. LLW continued this work until its completion and helped run the deconvolution methods. AD did the remaining analysis steps and package finalizing. ML and AD provided supervision of KP and LLW. CO initiated and coordinated the database and laboratory work. MW and SK acquired the laboratory data, supervised by AK and MC. AD drafted the manuscript. All authors revised the manuscript and gave their approval.

Ethics Statement

This article does not contain any studies with animals performed by any of the authors. All patients gave written consent. The study protocol conformed to the ethical guidelines of the Declaration of Helsinki, and the local ethics committee approved this study a priori (No. 9474_BO_K_2020).

Conflict of Interest Disclosures

MC reports personal fees from Abbvie, personal fees from Falk Foundation, personal fees from Gilead, personal fees from GlaxoSmithKline, personal fees from Jansen-Cilag, personal fees from Merck/MSD, personal fees from Novartis, personal fees from Roche, personal fees from Spring Bank Pharmaceuticals, and personal fees from Swedish Orphan Biovitrum, outside the submitted work. ML consults for mbiomics GmbH.

Other authors have nothing to disclose.

References

1. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018;67: 1031–1040.

2. Petitprez F, Sun C-M, Lacroix L, Sautès-Fridman C, de Reyniès A, Fridman WH. Quantitative Analyses of the Tumor Microenvironment Composition and Orientation in the Era of Precision Medicine. Frontiers in Oncology. 2018. doi:10.3389/fonc.2018.00390

3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63.

4. Merotto L, Zopoglou M, Zackl C, Finotello F. Next-generation deconvolution of transcriptomic data to investigate the tumor microenvironment. International Review of Cell and Molecular Biology. Academic Press; 2023.

5. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9: 5068.

6. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457.

7. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13: 86.

8. Hicks SC, Irizarry RA. methylCC: technology-independent estimation of cell type composition using differentially methylated regions. Genome Biol. 2019;20: 261.

9. Sturm G, Finotello F, Petitprez F, Zhang JD, Baumbach J, Fridman WH, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35: i436–i445.

10. Dietrich A, Merotto L, Pelz K, Eder B, Zackl C, Reinisch K, et al. Benchmarking second-generation methods for cell-type deconvolution of transcriptomic data. bioRxiv. 2024. p. 2024.06.10.598226. doi:10.1101/2024.06.10.598226

11. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics. 2017;18: 105.

12. De Ridder K, Che H, Leroy K, Thienpont B. Benchmarking of methods for DNA methylome deconvolution. Nat Commun. 2024;15: 4134.

13. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4: 2612.

14. Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, et al. A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. Epigenomics. 2018;10: 925–940.

15. Arneson D, Yang X, Wang K. MethylResolver-a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. Commun Biol. 2020;3: 422.

16. Patel H, Ewels P, Manning J, Garcia MU, Peltzer A, Hammarén R, et al. nf-core/rnaseq: nf-core/rnaseq v3.18.0 - Lithium Lynx. Zenodo; 2024. doi:10.5281/ZENODO.14537300

17. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38: 276–278.

18. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

19. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biol. 2019;20: 55.

20. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30: 1363–1369.

21. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7: e41361.

22. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012;13: 166–178.

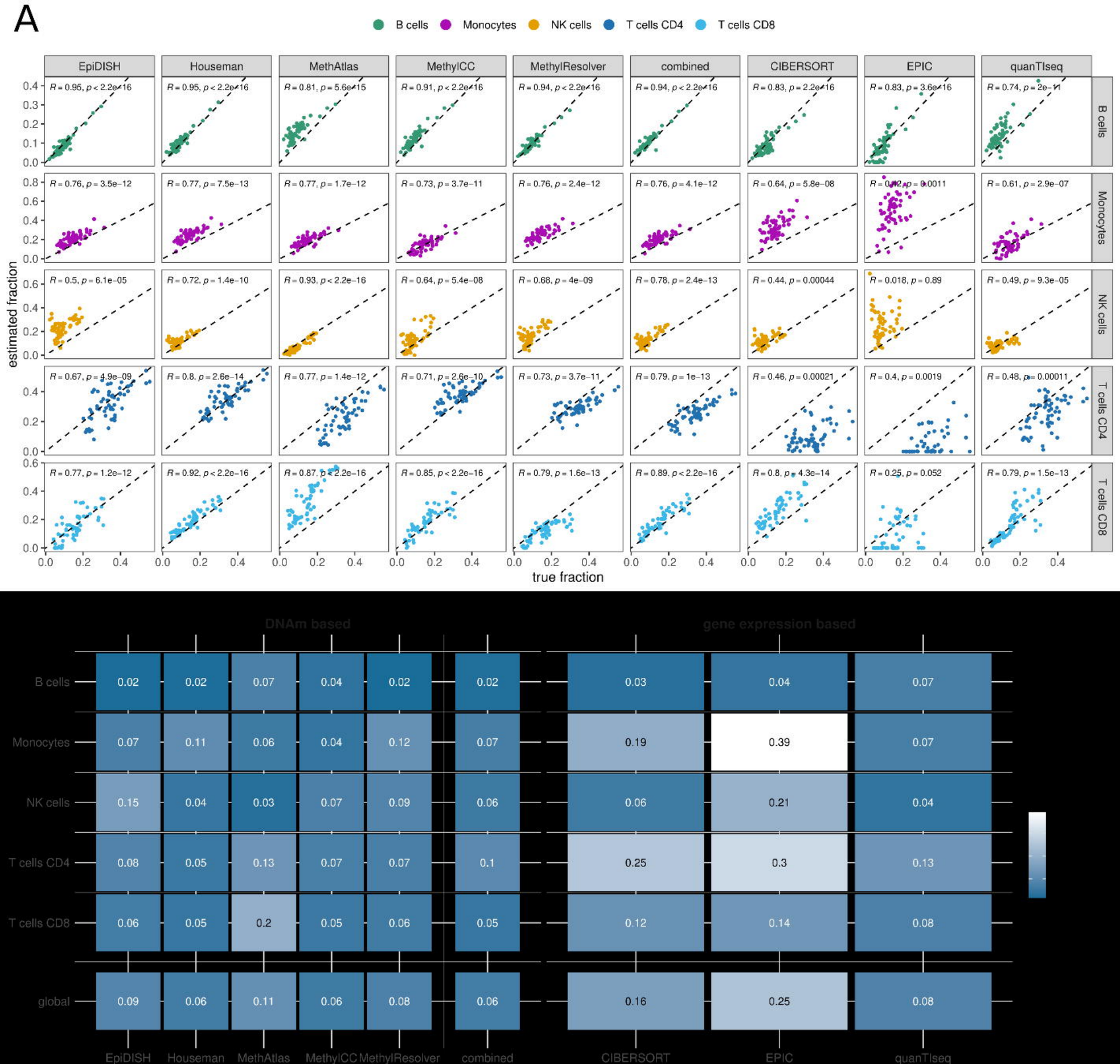
23. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25: 1841–1842.

24. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34: D590–8.

25. Finotello F, Mayer C, Plattner C, Laschober G, Rieder D, Hackl H, et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. Genome Med. 2019;11: 34.

26. Racle J, Gfeller D. EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data. Methods Mol Biol. 2020;2120: 233–248.

27. Gu Z. Complex heatmap visualization. Imeta. 2022;1: e43.


A
B cells
Monocytes
NK cells
T cells CD4
T cells CD8
EpiDISH
Houseman
MethAtlas
MethylCC
MethylResolver
combined
CIBERSORT
EPIC
quanTIseq
estimated fraction
true fraction
DNAm based
gene expression based
global